Supplementary Material of PaRPI

Supplementary Information 1: Competing methods

To demonstrate the effectiveness of our proposed model, we compare PaRPI against several deep learning models and machine learning algorithms as follows:

- HDRNet [1] (https://github.com/zhuhr213/HDRNet) is an end-to-end deep learning framework designed to predict dynamic RNA-binding protein (RBP) interactions across different cellular conditions. It integrates multi-source RNA data, including in vivo secondary structure and bio-language features, to capture both sequence and structural characteristics of RNA. By using hierarchical multi-scale residual networks (HMRN) and a deep protein-RNA binding predictor (DPRBP), HDRNet effectively models contextual dependencies and identifies key nucleotide tokens.
- PrismNet [2] (https://github.com/kuixu/PrismNet) is a recent work that introduced a deep learning approach using convolutional neural networks (CNN) to integrate RNA structure data from in vivo experiments with RBP binding information for accurate predictions of RBP binding sites. This method utilizes an "attention" mechanism to precisely pinpoint the nucleotides where RBPs bind. Remarkably, PrismNet is the first tool aimed at dynamic prediction tasks.
- PRIESSTESS [3] (https://github.com/kaitlin309/PRIESSTESS) presents a versatile RNA motif identification method that can scan and identify enriched RNA sequences and structural motifs. PRIESSTESS operates in two phases: the first generates a broad set of enriched motifs, including both sequence and structure information, while the second step aggregates these motifs into a single model, evaluating the importance of each motif.
- iDeep [4] (https://github.com/xypian1232/iDeep) introduces a novel hybrid approach combining convolutional neural networks (CNN) and deep belief networks for predicting RBP interaction sites and motifs. By transforming raw data into high-level abstractions using several layers of learning blocks, iDeep integrates shared features across different domains to enhance prediction accuracy.
- DMSK [5] (https://github.com/Rebecca3150/DMSK) is an innovative approach for identifying circRNA-RBP interaction sites. It combines multi-view deep learning, subspace learning, and multi-view classifiers, leveraging computationally predicted RNA secondary structures to improve identification accuracy.
- GraphProt [6] (https://github.com/dmaticzka/GraphProt) uses graph kernel methods to model the binding preferences of RBPs on RNA sequences. By incorporating both sequence and predicted RNA structural data, GraphProt effectively captures the binding preferences at RBP binding sites.
- DeepBind [7] (https://github.com/jisraeli/DeepBind) is a deep learning model based on CNN that predicts RBP binding sites using only RNA sequences.

Supplementary Information 2: Extended performance evaluation of PaRPI

We present the Receiver Operating Characteristic (ROC) and Precision-Recall (PRC) curves for two randomly selected datasets, as shown in Figure S1, to provide a visual comparison of model performance. The curves demonstrate that PaRPI achieves higher True Positive Rates (TPR), as well as improved Precision and Recall, compared to HDRNet and PrismNet. For instance, on the KHDRBS1_K562 dataset, PaRPI achieved an AUC of 0.98 and an AUPRC of 0.97, indicating slightly superior discriminative ability. These results further support that PaRPI is more effective at identifying true positive binding events, highlighting its enhanced sensitivity in RBP binding site prediction.

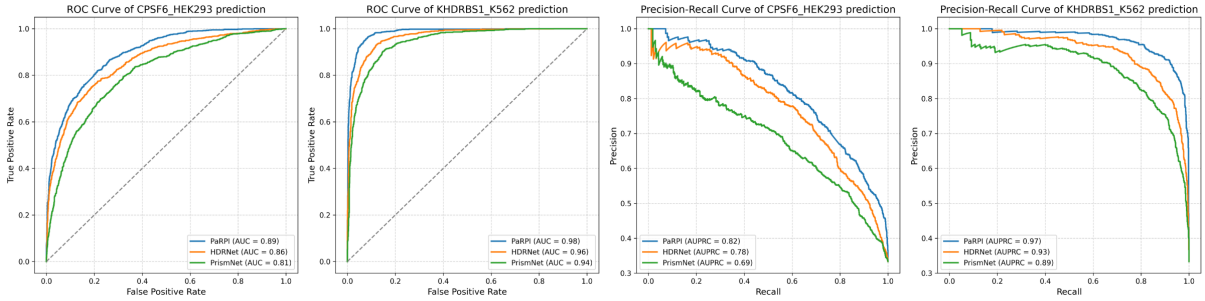

Figure S1. ROC and PRC curves for PaRPI and other methods on the CPSF6_HEK293 and KHDRBS1_K562 datasets, prior to redundancy reduction.

Supplementary Information 3: Evaluating the performance of Various pre-trained RNA models

We conducted a comparative analysis between our proposed K-mer encoding approach and several state-of-the-art pre-trained RNA models, including BERT [8], RNABERT [9], RNA-FM [10] and RiNALMo [11]. All models were trained on the integrated dataset derived from the Hela cell line, ensuring a fair comparison. As shown in Table S1, although these pre-trained models are capable of extracting meaningful sequence representations, our K-mer-based strategy achieved slightly better performance.

This improvement may be attributed to the better alignment between our encoding strategy and the specific characteristics of the RNA–protein interaction prediction task. The K-mer representation explicitly captures local sequence composition patterns, which are known to play a critical role in determining RNA–protein binding specificity. In contrast, many pre-trained RNA language models are trained on large-scale general RNA datasets (e.g., from Rfam or GenBank), which may not emphasize these task-specific motifs to the same extent.

Table S1: Performance comparison of PaRPI on evaluation metrics.

Model	AUC	ACC	F1	MCC
K-mer+BERT	**0.938**	**0.821**	**0.782**	**0.652**
BERT	0.925	0.808	0.764	0.637
RNABERT	0.930	0.793	0.749	0.615
RNA-FM	0.921	0.818	0.765	0.638
RiNALMo	0.915	0.778	0.737	0.597

Supplementary Information 4: Calculation formulas of the icSHAPE method for analyzing in vivo RNA secondary structures

We employed icSHAPE [12] to characterize the in vivo RNA secondary structure under different cellular conditions.The icSHAPE structure score **R** is computed based on reverse transcription (RT) signal counts and polymeric sequencing coverage (base density, BD) at individual nucleotide positions. During sequencing, each mapped read contributes an RT value for the first base upstream of the mapping start point and a BD value for all bases covered. For each sequence, a sliding window of size **wSize** is initialized at the 5' end and moves in the 3' direction with window step **wStep**. The RT and BD values from two replicates of the DMSO and NAI libraries are combined by direct addition as follows:

$$r_i^C = r_i^C 1 + r_i^C 2$$

$$r_i^T = r_i^T 1 + r_i^T 2$$

$$b_i^C = b_i^C 1 + b_i^C 2$$

$$1 \leq i \leq wSize$$

where r_i^C is the RT signal from the DMSO library, r_i^T is the RT signal from the NAI library, and b_i^C is the BD for the DMSO library. The values are then normalized by dividing by the average of the 90th to 95th percentile values, as follows:

$$r_i^C = \frac{r_i^C}{r_{q95}^C}$$

$$r_i^T = \frac{r_i^T}{r_{q95}^T}$$

$$b_i^C = \frac{b_i^C}{b_{q95}^C}$$

where r_{q95}^C, r_{q95}^T, and b_{q95}^C are the normalization factors calculated for each replicate. Based on this, the enrichment signal is computed as:

$$e_i = \frac{r_i^T - \alpha \cdot r_i^1}{b_i^C}$$

where α is a subtraction factor to account for background noise from the DMSO sample that might influence the NAI sample signal. Finally, the normalized icSHAPE score R is computed as follows:

$$R_i = \begin{cases} \min\left(1, \max\left(0, \frac{e_i - e_{q95}}{e_{q95} - e_{q5}}\right)\right), & b_i^c \geq 200 \\ \text{NULL}, & b_i^c < 200 \end{cases}$$

where e_{q5} represents the 5% percentile of the signal, and e_{q95} represents the 95% percentile. Notably, only nucleotides with coverage exceeding 200x are considered to ensure sequencing quality control. Consequently, the RNA sequences are analyzed using the icSHAPE technique, which produces equal-length vectors of numerical values for each RNA fragment. These vectors encode the secondary structural features of RNA and serve as input for identifying RNA-binding events under diverse cellular conditions.

Supplementary Information 5: Hyperparameter tuning and model evaluation

To evaluate the model's performance and stability under different hyperparameter settings, we randomly selected 10 sets of hyperparameters and trained the model on the validation set of HEK293 cell line dataset. Each configuration was evaluated based on multiple metrics including AUC, ACC, F1 score, and MCC, and the optimal combination was selected according to the highest overall performance across these metrics, as shown in Table S2.

Table S2: The performance of PaRPI using different hyperparameters

Hidden Units	h	Dropout	L_1	L_2	lr	AUC	ACC	F1	MCC
256	8	0.4	1	4	1e-4	**0.871**	0.758	**0.717**	**0.567**
128	8	0.3	1	4	1e-4	0.866	0.774	0.706	0.560
64	8	0.2	1	4	1e-4	0.852	0.704	0.695	0.545
256	4	0.4	1	4	1e-4	0.863	0.723	0.705	0.558
256	2	0.4	1	4	1e-4	0.862	0.772	0.707	0.560
256	8	0.4	1	5	1e-4	0.858	**0.797**	0.699	0.548
256	8	0.4	1	3	3e-4	0.835	0.749	0.672	0.504
256	8	0.4	1	2	2e-4	0.830	0.758	0.666	0.481
256	8	0.4	2	4	1e-4	0.813	0.746	0.649	0.455
256	8	0.4	3	4	1e-4	0.800	0.640	0.608	0.362

Note: Hidden units represent the dimension of the hidden layer, h denotes the number of attention heads, and Dropout refers to the dropout rate applied in each layer. L_1 and L_2 represent the number of Transformer and GraphSAGE layers, respectively. lr is the learning rate. AUC, ACC, F1, and MCC are evaluation metrics used to assess model performance.

Supplementary References

[1] Haoran Zhu and et al. Dynamic characterization and interpretation for protein-rna interactions across diverse cellular conditions using HDRNet. *Nature Communications*, 14(1):6824, 10 2023.

[2] Lei Sun and et al. Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research*, 31(5):495–516, 2021.

[3] Kaitlin U Laverty and et al. Priesstess: interpretable, high-performing models of the sequence and structure preferences of rna-binding proteins. *Nucleic Acids Research*, 50(19):e111–e111, Oct 2022.

[4] Xiaoyong Pan and Hong-Bin Shen. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18:1–14, 2017.

[5] Hui Li and et al. circrna-binding protein site prediction based on multi-view deep learning, subspace learning and multi-view classifier. *Briefings in Bioinformatics*, 23(1):bbab394, 2022.

[6] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. Graphprot: modeling binding preferences of rna-binding proteins. *Genome biology*, 15:1–18, 2014.

[7] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[8] Jacob Devlin and et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[9] Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4(1):lqac012, 02 2022.

[10] Jiayang Chen and et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[11] Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*, 2024.

[12] Ryan A Flynn and et al. Transcriptome-wide interrogation of rna secondary structure in living cells with icshape. *Nature Protocols*, page 273–290, Feb 2016.